ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 23-03-01
j.inst.Korean.electr.electron.eng.Vol.27,No.3,213~219,September 2023
https : //doi.org/10.7471/ikeee.2023.27.3.213

GRU 기반 단축 URL 판별 기법을 적용한 하이브리드 피싱 사이트 탐지 시스템

Hybrid phishing site detection system with GRU-based shortened URL determination technique

김 해 수*, 김 미 희*★

Hae-Soo Kim*, Mi-Hui Kim*★

Abstract

According to statistics from the National Police Agency, smishing crimes using texts or messengers have increased dramatically since COVID-19. In addition, most of the cases of impersonation of public institutions reported to agency were related to vaccination and reward, and many methods were used to trick people into clicking on fake URLs (Uniform Resource Locators). When detecting them, URL-based detection methods cannot detect them properly if the information of the URL is hidden, and content-based detection methods are slow and use a lot of resources. In this paper, we propose a system for URL-based detection using transformer for regular URLs and content-based detection using XGBoost for shortened URLs through the process of determining shortened URLs using GRU(Gated Recurrent Units). The F1-Score of the proposed detection system was 94.86, and its average processing time was 5.4 seconds.

요 약

경찰청 통계자료에 따르면 코로나19 이후 문자 또는 메신저를 이용한 스미싱(Smishing) 범죄가 급증하였다. 또한 정부 기관에 접수된 공공기관 사칭 건수의 대부분이 백신접종 및 보상 관련하여 가짜 URL(Uniform Resource Locator)을 클릭하도록 유도하는 수법이 다수 사용되었다. 주로 URL의 정보를 숨긴 단축 URL을 사용하며 탐지할 때 URL 기반 탐지방법은 URL의 정보를 숨기면 제대로 탐지할 수 없고, 콘텐츠 기반 탐지 방법은 탐지 속도가 느리고 많은 자원을 사용한다. 이에 본 논문에서는 GRU(Gated Recurrent Units)를 이용한 단축 URL을 판별하는 과정을 통해 일반 URL일 때 transformer를 통한 URL 기반 탐지, 단축 URL일 때 XGBoost를 이용한 콘텐츠 기반 탐지하는 시스템을 제안한다. 제안한 탐지 시스템의 F1-Score는 94.86이었고, 처리시간은 평균 5.4초가 소요되었다.

Key words : Phishing, Artificial Intelligence, Deep learning, Security, Phishing Detection

* School. of Computer Engineering & Applied Mathematics, Computer System Institute, Hankyong National University
★ Corresponding author
E-mail : mhkim@hknu.ac.kr, Tel : +82 31-670-5167
※ Acknowledgment
This work was supported by a research grant from Hankyong National University for an academic exchange program in 2023.
Manuscript received Jun. 3, 2023; revised Jun. 25, 2023; accepted Jul. 20, 2023.

Ⅰ. 서론

경찰청 통계자료에 따르면 코로나19 이후 사회적 이슈를 이용한 사이버 범죄가 증가하는 추세이고 그 중 문자 또는 메신저를 이용한 스미싱(Smishing) 범죄가 2019년도에 207건, 2020년도에 822건, 2021년도에 1,336건이 경찰에 접수되어 3년간 건수가 급증하였다. 스미싱이란 문자메시지를 뜻하는 SMS(Short Message Service)와 피싱(Phishing)의 합성어로 문자메시지를 통해 직, 간접적인 개인정보 탈취를 위한 행위를 의미한다. 또한 KISA(Korea Internet & Security Agency)에 접수된 공공기관 사칭 건수가 2019년도에 36건, 2020년도에 12,208건, 2021년도에 16,513건으로 백신접종 및 보상 관련하여 가짜 URL(Uniform Resource Locator)을 클릭하도록 유도하는 수법이 다수 사용되었다[1].

표 1은 스미싱(a)과 정상 문자(b)이다. 표 1의 URL과 같이 정부 혹은 공공기관의 URL이 아닌 단축 URL의 경우 수신자가 직접 웹 사이트에 접속하지 않으면 스미싱 유무를 구별할 수 없다.

Table 1. Example of Smishing[1] a) Smishing message b) Legitimate message.

표 1. 스미싱 예시[1] a) 스미싱 메시지 b) 정상 메시지

	Text	Type
a)	[질병관리청] 8/5 코로나19 백신 예약 인증 본인확인 https://ya.mba/3Pm	Smishing message
b)	[이천시청] ▶12월 3일 확진자 145명 발생 ▶동절기 재유행 대비 코로나 2가백신 추가접종 필요. 접종안내 ☞(https://url.kr/f1z9dj)	Legitimate message

이러한 피싱 사이트를 빠르게 탐지하기 위해 URL을 통해 www.naver.com, www.navers.com과 같이 기존에 알려진 유명 웹 사이트와 유사한 도메인을 확인하여 알려진 정보를 기반으로 피싱 사이트 여부를 판별하거나 콜백, 파라미터 등과 같은 정보들을 통해 판별하는 URL 기반 탐지 방법이 있다. 그러나 이러한 탐지 방법은 URL의 정보를 숨기면 탐지할 수 없다는 단점이 있다. 그래서 웹 사이트의 외부 하이퍼 링크, iframe과 같은 콘텐츠 정보를 추출하여 판별하는 콘텐츠 기반 탐지도 가능하지만, 탐지 속도가 느리고 많은 자원을 사용하는 단점이 있다. 따라서 적절한 기준을 통해 URL과 콘텐츠 기반 탐지를 할 수 있는 시스템이 필요하다. 이에 [2]에서는 표 1과 같은 단축 URL이 URL의 정보를 숨기기 위해 사용된다는 점을 이용해 단축 URL을 기준으로 단축 URL일 때는 콘텐츠 기반 탐지 기법을 이용하고 아닐 때는 URL 기반 탐지 기법을 이용하여 피싱 URL을 판별 기법을 제안하였다. 선행 연구인 [2]에서 제안한 단축 URL 판별 방식은 많은 단축 URL 서비스의 도메인을 데이터베이스화할 수 없으므로 본 논문에서는 단축 URL과 일반 URL 도메인의 차이를 딥러닝을 통해 학습하여 판별하는 방식으로 개선하였다.

본 논문은 2장에서 피싱 사이트 탐지에 관한 관련 연구를 소개하고, 3장에서는 제안하는 시스템을 설명한다. 4장에서는 제안 시스템을 실험한 결과를 분석하고 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

1. 피싱 URL 탐지 기법

[3]의 연구에서는 URL 기반 피싱 사이트 탐지에 관해 연구했으며 URL에서 추출할 수 있는 URL의 길이, 도메인의 문자 개수, 도메인 길이, 단축 URL의 여부 등의 특성을 SVM(Support Vector Machine), Decision Tree, Random Forest 알고리즘을 이용해서 훈련을 진행하여 비교하였고 그 결과 Random Forest 알고리즘이 가장 높은 정확도를 달성함을 보였다.

[4]의 연구에서는 콘텐츠 기반 피싱 사이트 탐지 방법을 연구했다. 58개의 특성을 추출하여 특성의 유무, 개수, 길이를 데이터 세트로 구성하여 Naive Bayes, Random Forest, SVM, Logistic Regression, K-NN(K-Nearest Neighbor), Decision Tree, MLP(Multilayer Perceptron), XGBoost 알고리즘을 이용하여 훈련을 진행하였고 XGBoost가 가장 높은 정확도를 보였다.

2. URL 및 콘텐츠 기반 탐지 기법 결합

[5]의 연구에서는 URL에서 추출한 도메인 네임, '.'의 개수, IP주소의 유무, '@'의 유무, URL의 길이, '/'으로 나뉜 서브 페이지 혹은 서브 폴더의 개수, 'http/https'와 같은 프로토콜의 유무, https 프로토콜의 여부, 프로토콜 표시를 제외한 '//'의 유무, URL 내부에서의 단축 URL의 존재 여부 등과 같은 정보들과 크롤러를 통해 웹 사이트의 소스 코드를 추출해 하이퍼링크의 유무, 내부 링크로 이동하는 하이퍼링크의 비율, 외부 링크로 이동하는 하이퍼링크의 비율 등을 하나의 데이터로 결합하여 총 25개의 특성을 이용하여 예측하는 기법을 제안하였다. 해당 기법은 URL만으로 탐지를 할 수 있거나 URL의 정보가 탐지에 영향을 주지 않는 경우 URL 기반 혹

은 콘텐츠 기반 단일로 탐지하지 않고 두 기법 모두 이용한다는 단점이 있다.

[6]에서는 GAN(Generative Adversarial Network)과 CNN(Convolution Neural Network) 모델을 결합하여 URL 기반 탐지 후 정상 웹 사이트일 때 DNN 모델을 통해 콘텐츠 기반 탐지를 수행하여 URL, 콘텐츠 기반 2단계 탐지 기법을 제안하였다. 피싱 사이트와 정상 사이트의 비율이 반반이라고 가정할 때 정상 사이트 대부분은 콘텐츠 기반 기법으로 탐지하게 될 것이고 오탐 된 일부 피싱 사이트 또한 콘텐츠 기반 기법으로 탐지하게 된다. 단축 URL이 데이터에 존재한다면 정상 사이트와 URL만으로 정보를 알 수 없는 단축 URL이 이중으로 URL과 콘텐츠 기반 탐지하는 단점을 갖고 있다.

Ⅲ. 제안 시스템

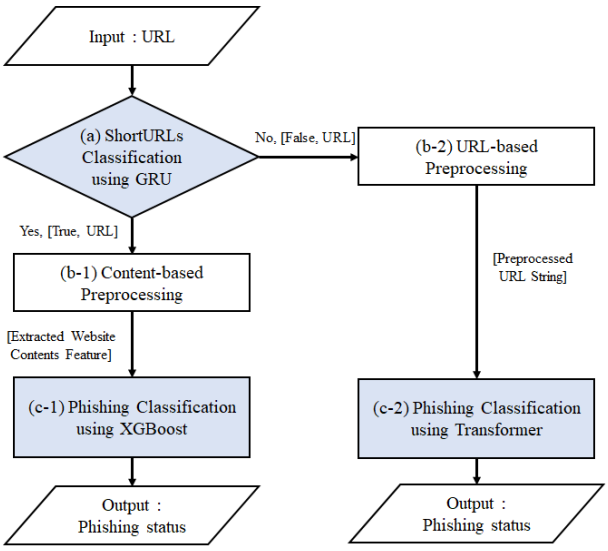

Fig. 1. Flowchart of Proposal System.
그림 1. 제안 시스템의 흐름도

본 장에서는 제안하는 시스템을 설명한다. 그림 1은 제안하는 탐지 시스템의 순서도이다. 크게 3단계로 나뉜다; (a)단축 URL 판별; (b) 피싱 사이트 탐지법(URL 기반 탐지, 콘텐츠 기반 탐지)에 맞게 전처리; (c) 피싱 사이트 탐지

(a, Short URLs Classification using GRU)는 입력인 URL이 단축 URL의 여부를 판별하는 단계이며 단축 URL일 때 (b-1)을 진행하여 콘텐츠 기반 탐지를 위한 전처리 후 추출된 웹 사이트의 콘텐츠 특징들을 통해 (c-1, Phishing Classification using XGBoost)에서 피싱 사이트 여부를 판별한다. (a)의 과정에서 입력된 URL이 단축 URL이 아닐 때 (b-2)에서 URL 문자열을 다음 단계인 (c-2, Phishing Classification using Transformer)의 입력에 맞게 전처리를 진행한다. 이후 다음 단계에서 전처리 된 URL 문자열을 통해 피싱 사이트 여부를 판별한다.

1. 단축 URL 판별

판별에 사용되는 모델은 GRU(Gated Recurrent Unit) 모델로 기존 LSTM(Long Short-Term Memory) [7]의 구조를 간소화 시킨 모델이다[8].

단축 URL의 길이가 짧다는 점을 통해 GRU 모델을 사용하여 LSTM보다 빠른 속도 및 높은 정확도를 기대할 수 있다.

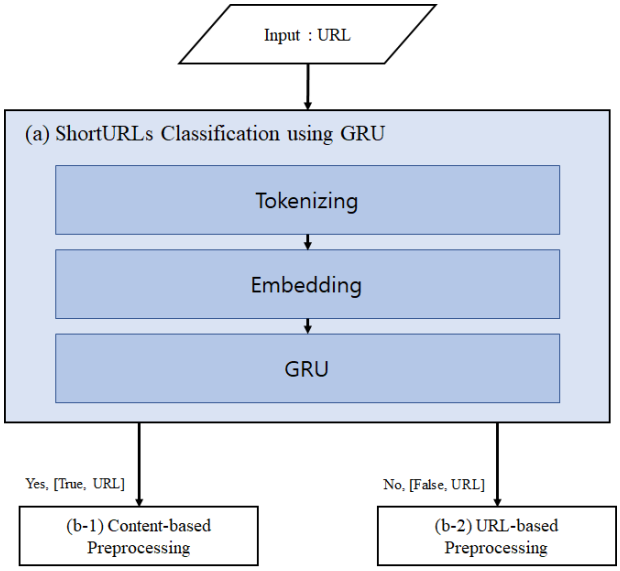

Fig. 2. Components of Short-URLs Classification Model using GRU.
그림 2. GRU를 이용한 단축 URL 분류 모델의 구성요소

그림 2는 단축 URL 판별을 위한 GRU 모델의 구성요소이다. Tokenizing은 데이터로 단어 집합을 생성하고, 생성된 단어 집합으로 임의의 문장을 의미 있는 단어 단위로 나누는 과정이다. Embedding은 단어들을 밀집 벡터화하는 과정이다. 밀집 벡터란 0과 1로 이루어져 고차원의 데이터가 생성되는 원-핫 벡터와는 달리 모든 데이터가 실수로 이루어진 저차원 데이터로 생성되어 모든 데이터를 지정한 차원으로 축약 혹은 확장할 수 있어 같은 크기의 저차원의 데이터를 생성할 수 있다. 이를 통해 다양한 문자로 구성된 URL을 수치화할 수 있게 한다. Embedding 과정으로 생성된 데이터를 GRU 계층을 통해 단축 URL 여부를 판별한다.

Table 2. Structure of GRU model.
표 2. GRU 모델 구조

Layer	Parameters	Values	Output
Embedding Layer	input_dim	number of train dataset	1000×100
	output_dim	100	
	input_length	1000	
GRU Layer_1	units	150	100×150
	return sequences	true	
	activation	tanh	
Dropout Layer	rate	0.5	100×150
GRU Layer_2	units	150	150
	activation	tanh	
Dense	units	1	1
	activation	sigmoid	

표 2는 본 논문에서 사용한 GRU 모델의 네트워크 구조와 하이퍼 파라미터이다.

2. URL 기반 탐지

관련 연구에서 소개한 연구들[3-4]에서 URL 기반 탐지에 쓰인 방식은 URL에서 정보를 추출 후 수치형 데이터로 전처리하여 훈련하는 방식이다. 실시간으로 피싱 URL을 탐지할 때 URL 전체를 확인해 사전에 정의된 정보들을 추출하여 머신러닝 알고리즘에 훈련하는 과정을 진행하면 탐지과정이 이중으로 진행되어 판별 과정뿐만 아니라 데이터를 추출하는 과정 또한 Transformer를 이용해 해결한다. Transformer는 Attention 기법만으로 구성되어 RNN(Recurrent Neural Network) [9]의 단점을 개선한 모델이다[10]. Transformer는 순차적으로 입력 받는 것이 아닌 한 번에 데이터를 입력받아 병렬로 처리하는 것이 가장 큰 특징이며 URL 정보를 추출하여 훈련하지 않고 URL 전체 문자열을 입력으로 하여 판별할 수 있게 된다.

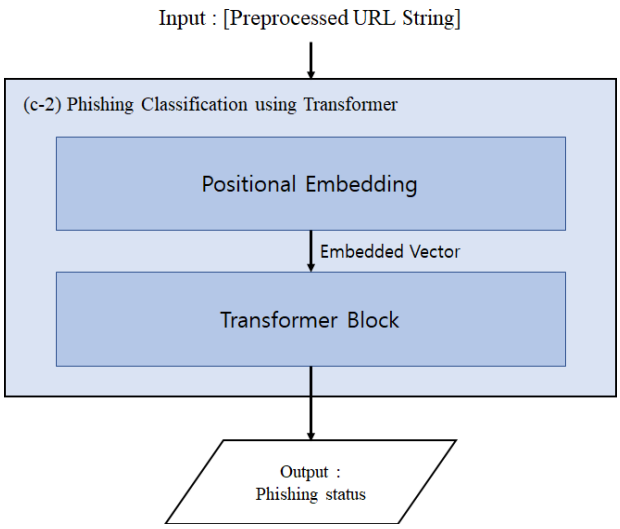

Fig. 3. Components of Phishing Classification Model using Transformer.
그림 3. Transformer를 이용한 단축 URL 분류 모델의 구성요소

그림 3은 Transformer 모델의 구성요소이다. Transformer는 입력되는 문자열을 구성하는 각 단어의 순서 정보는 학습하지 않는다. 따라서 각 단어의 순서정보를 입력에 추가해야 하며 해당 과정이 Positional Embedding이다. Embedding Layer에서 각 단어의 위치마다 고유한 값을 만들어 내는 적절한 함수를 이용해, 문장에서 각 단어의 위치를 설명하는 위치 임베딩 벡터를 만든 뒤, 단어 임베딩 벡터에 더하는 과정을 통해 임베딩 된 벡터(Embedded vector) 배열을 반환한다. Transformer Block은 트랜스포머 모델을 구성하는 계층의 집합이다. 트랜스포머 블록에서 쿼리(Query), 키(Key), 값(Value)을 통해 어텐션을 계산하고 계산된 어텐션 값을 순방향 신경망(feed forward network)과 분류 층을 통해 결과를 출력한다. 단어의 위치정보를 전처리 된 URL에 추가하고 Transformer Block의 모델을 통해 피싱 사이트 여부를 판별한다.

표 3은 본 논문에서 사용된 모델의 네트워크 구조와 구성하고 있는 하이퍼 파라미터이다.

Table 3. Structure of Transformer model.
표 3. Transformer 모델 구조

Layer	Parameters	Values	Output
Embedding Layer	word_embeddings	30000×128	128
	position_embeddings	512×128	
	token_type_embeddings	128×128	
	LayerNorm	128	
Transformer Block	embedding_hidden_mapping_in	in=128	768
		out=768	
	full_layer_layer_norm	in=(768,)	
		eps=1e-12	
	attention	query	
		key	
		value	
	feed_forward_network	in=768	
		out=768	
	activation	NewGELUActivation	
classifier	in_features	768	2
	out_features	2	

3. 콘텐츠 기반 탐지

콘텐츠 기반 탐지는 웹페이지에서 특징들을 추출하는 과정이 필수적이며 제안 기법 전체에서 판별하는 속도 대부분을 해당 부분이 차지한다. 따라서 모델이 판별하는 과정에서는 정확도를 확보하고 소요 시간을 단축해야 한다.

[4]의 결과에 따라 높은 정확도와 짧은 실행 시간을 보여준 XGBoost를 사용한다. XGBoost는 Boosting 기법을 사용하는 알고리즘 중에 하나로, Boosting은 머신러닝의 앙상블 기법이며 여러 개의 약한 학습기를 결합하여 예측 또는 분류의 성능을 높이는 기법이다. Boosting의 대표적인 모델로 GBM(Gradient Boosting Algorithm)[11]이 있다. XGBoost는 Gradient Boosting 알고리즘을 병렬처리가 가능하도록 구현한 라이브러리이다[12].

표 4는 본 논문의 콘텐츠 기반 탐지에 사용될 특성이다.

각 URL에서 해당 특성들을 추출해서 XGBoost 알고리즘을 통해 피싱 사이트 여부를 판별한다.

Table 4. Features of Web-page contents[13].
표 4. 웹 페이지 콘텐츠의 특징[13]

#	Features of Web-page contents
1	Number of hyperlinks present in a website
2	Internal hyperlinks ratio
3	External hyperlinks ratio
4	Number of null hyperlinks
5	External CSS
6	Internal redirection
7	External redirection
8	Generates internal errors
9	Generates external errors
10	Having login form link
11	Having external favicon
12	Submitting to email
13	Percentile of internal media
14	Percentile of external media
15	Check for empty title
16	Percentile of safe anchor
17	Percentile of internal links
18	Server Form Handler
19	iframe Redirection
20	On mouse action
21	Pop up window
22	Right_click action
23	Domain in page title
24	Domain after copyright logo

Ⅳ. 성능 평가

본 장에서는 3장에서 소개한 시스템의 성능을 측정하기 위한 실험 환경 및 데이터 세트를 소개하고 실험 결과를 분석한다.

1. 실험 환경

실험 환경은 Intel Xeon(R) Silver 4215R CPU @ 3.20GHz CPU, 256GB RAM, NVIDIA RTX A6000 GPU 환경에서 진행한다.

가. 데이터 세트[13]

사용한 데이터는 Mendeley Data에서 공개된 “Web page phishing detection”이라는 데이터를 사용하였고 11,430개의 데이터를 사용하였으며 피싱 URL과 정상 URL은 각각 5,715개이고, 단축 URL의 개수는 1,411개다. 단축 URL 중 피싱 URL은 905개이고 정상 URL은 506개이다. 훈련, 검증 및 테스트 데이터는 60:20:20 비율로 나눠 실험했다.

나. 평가 방법

성능 평가 방법은 GRU 모델의 단축 URL 판별 정확도 및 F1-Score와 URL 기반, 콘텐츠 기반 탐지에 대한 정확도와 F1-Score, 특성 추출부터 탐지까지의 소요 시간을 측정한다. F1-Score는 정밀도와 재현율의 조화평균으로 정밀도는 모델이 True로 예측한 값이 실제 True인 비율이고 재현율은 실제 True 중에 모델이 True로 예측한 비율이다. 따라서 F1-Score는 한쪽에 치우쳐진 데이터를 평가할 때 효과적인 지표이다.

2. 실험 결과 분석

Table 5. Shortened URL Discrimination Accuracy
표 5. 단축 URL 판별 정확도

	Shortened URL Discrimination Accuracy
Accuracy	98.32%
F1-Score	93.22%

표 5는 단축 URL 판별 정확도를 보여준다. 정확도는 98.32%를 달성하였으나 f1-score는 93.22%로 5.1% 차이가 난다. 그림 4는 단축 URL 분류의 결과를 Confusion Matrix로 표현한 그림이다. 왼쪽(True label)은 실제 데이터의 라벨이고 아래(Predicted label)는 분류 모델이 분류한 결과 라벨이다. 총 2,858개(네 군락의 총합)의 테

스트 데이터 중 일반 URL 2,483개(True label의 normal 행의 총합), 단축 URL 375개(True label의 short 행의 총합)에서 일반 URL의 오탐 데이터는 2,483개 중 2개였으나 단축 URL의 오탐 데이터는 375개 중 45개이다.

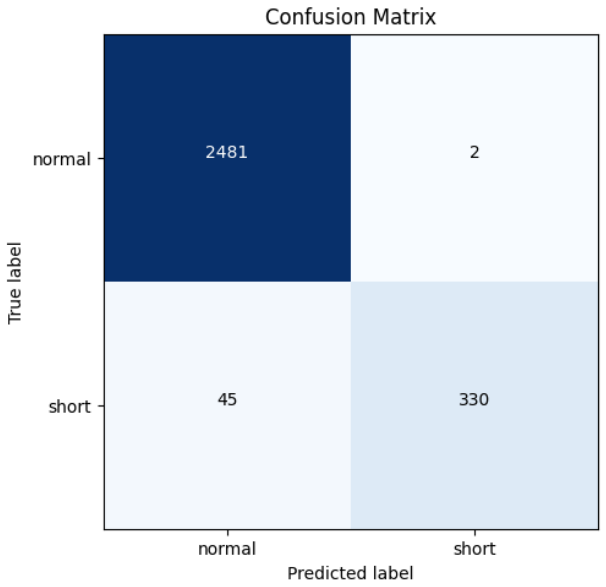

Fig. 4. Confusion Matrix of Shortened URL Discrimination.
그림 4. 단축 URL 판별 Confusion Matrix

표 6은 제안 시스템의 정확도와 f1-score를 측정한 표이다.

URL 기반 탐지에서 정확도는 89.18%, f1-score는 89.39%를 보였고 콘텐츠 기반 탐지에서 정확도는 95.59%, f1-score는 95.56%를 보였다. 제안 시스템은 94.85%의 정확도와 94.86%의 f1-score를 보였다. URL 기반 탐지에 비해 5.67%의 정확도와 5.47%의 f1-score의 향상했고 콘텐츠 기반 탐지에서는 정확도가 0.74%, f1-score가 0.7% 하락하였다.

Table 6. Overall System Accuracy.
표 6. 전체 시스템 정확도

	URL based Detection[3]	Contents based Detection[4]	Proposal System
Accuracy	89.18%	95.59%	94.85%
F1-Score	89.39%	95.56%	94.86%

표 7은 제안 시스템이 탐지에 소요한 시간을 보여주는 표이다. 사용된 데이터는 3개의 활성화되어있는 URL을 이용하였고 2개의 단축 URL과 1개의 일반 URL을 이용하였다. 3개의 URL 모두 URL 기반, 콘텐츠 기반 탐지를 각각 수행하고 제안 시스템을 이용해 탐지하는 과정 동안 소요되는 평균 시간을 측정하였다. URL 기반 탐지에서 URL 추출 평균 시간은 351ms, URL 기반 탐지 시간은 7ms가 소요되었고 전체 소요 시간은 358ms이다. 콘텐츠 기반에서 웹 사이트의 콘텐츠 추출 평균 시간은 8,342ms, 콘텐츠 기반 탐지 평균 시간은 61ms가 소요되었고 전체 소요 시간은 8,403ms이다. 제안 시스템은 1개의 URL에 대한 URL 탐지 시간은 227ms, 2개의 URL에 대한 평균 콘텐츠 추출 시간은 7,993ms, 평균 탐지 시간은 59.5ms 소요되었다. 3개의 URL에 대한 제안 시스템의 전체 평균 소요 시간은 5,404ms가 소요되었다.

Table 7. Detection time(ms).
표 7. 탐지 시간(밀리초)

	URL based Detection[3]	Contents based Detection[4]	Proposal System
URL Extraction	351	-	-
URL based Detection	7	-	227
Contents Extraction	-	8,342	7,993
Contents based Detection	-	61	59.5
Average (Extraction + Detection)	358	8,403	5,404

3. 정확도와 전체 소요 시간 정리

일반 URL의 정보를 숨긴 단축 URL은 URL 기반 탐지에서 정확도를 하락시키는 원인이다. 또한 콘텐츠 기반 탐지는 정확도를 높일 수 있지만 탐지에 많은 시간이 소요된다. 본 논문은 단축 URL 판별 과정을 통해 일반 URL일 때 URL 기반, 단축 URL일 때 콘텐츠 기반 탐지를 진행하는 하이브리드 시스템을 제안하여 정확도 측면에서 4.2장의 표 6에서 제안 시스템이 URL 기반 탐지보다 5.67%, 1.06배 상승했고, 콘텐츠 기반 탐지보다 0.74%, 1.007배 하락했다. 그러나 표 7에서 제안 시스템이 콘텐츠 기반 평균 소요 시간보다 약 3,000ms로 1.55배 단축하여 콘텐츠 기반 탐지 기법보다 정확도 대비 짧은 소요 시간을 확보하였다.

V. 결론

본 논문에서는 단축 URL을 기준으로 일반 URL일 때는 URL 기반 탐지, 단축 URL일 때는 콘텐츠 기반 탐지하는 URL-콘텐츠 기반 하이브리드 탐지 시스템을 제안하였다.

실험 결과 GRU 모델은 단축 URL 탐지에서 좋은 성능을 보였다. 또한 제안 시스템에서 사용된 Transformer

와 XGBoost에서 탐지할 때 걸리는 시간은 평균 7ms에서 1ms 사이이지만 웹 사이트 콘텐츠를 추출하는 시간이 걸리는 시간 대부분임을 확인하였다.

향후 연구에서 콘텐츠 기반 탐지에서 피싱 사이트를 결정하는 주요 특성을 찾고 추출하는 시간을 최소화하는 방안을 연구하고자 한다.

References

[1] Korean National Police Agency, "Cyber Criminal Trend for 2022," https://www.police.go.kr/user/bbs/BD_selectBbs.do?q_bbsCode=1001&q_bbscttSn=20220519141449594.

[2] Y. Kim, H. Kim and M. Kim, "Short URLs Verification Approach for Phishing Site Detection Improvement," *Proceedings of the Annual Conference of Korea Information Processing Society Conference 2022 (ACK 2022)*, pp.80-81, 2022.

[3] J. Lamas Pineiro and L. Wong Portillo, "Web architecture for URL-based phishing detection based on Random Forest, Classification Trees, and Support Vector Machine," *Inteligencia Artificial*, vol.25, no.69, pp.107-121, 2022.
DOI: 10.4114/intartif.vol25iss69pp107-121

[4] U. Ozker and O. K. Sahingoz, "Content Based Phishing Detection with Machine Learning," *2020 International Conference on Electrical Engineering (ICEE)*, pp.1-6, 2020.
DOI: 10.1109/ICEE49691.2020.9249892

[5] Das Guptta, S., Shahriar, K. T., Alqahtani, H. et al., "Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," *Annals of Data Science*, 2022.
DOI: 10.1007/s40745-022-00379-8

[6] Korkmaz, M. ., Kocyigit, E. ., Sahingoz, O. K., & Diri, B., "A Hybrid Phishing Detection System Using Deep Learning-based URL and Content Analysis," *Elektronika Ir Elektrotechnika*, vol.28, no.5, pp.80-89, 2022. DOI: 10.5755/j02.eie.31197

[7] S. Hochreiter and J. Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, vol.9, no.8, pp.1735-1780, 1997.

[8] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014. [online] Available: http://arxiv.org/abs/1412.3555.

[9] Jordan. M. I, "Serial order: a parallel distributed processing approach," *Tech. rep. ICS 8604,* 1986.
DOI: 10.1016/S0166-4115(97)80111-2

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," *Advances in neural information processing systems 30*, 2017. DOI: 10.48550/arXiv.1706.03762

[11] J. H. Friedman, "Greedy function approximation: A gradient boostingmachine," *Ann. Statist*, vol.29, no.5, pp.1189-1232, 2001.
DOI: 10.1214/aos/1013203451

[12] dmlc XGBoost, "XGBoost Documentation," https://xgboost.readthedocs.io/en/stable/

[13] A. Hannousse, S. Yahiouche, "Web page phishing detection," https://data.mendeley.com/datasets/c2gw7fy2j4/3

BIOGRAPHY

Hae-Soo Kim (Member)

2022 : BS degree in Computer Science and Engineering, Hankyong National University
2022~present : MS student in School of Computer Engineering & Applied Mathematics, Hankyong National University

Mi-Hui Kim (Member)

1997 : BS degree in Computer Science and Engineering, Ewha Womans University.
1999 : MS degree in Computer Science and Engineering, Ewha Womans University.
1999~2003 : Researchers at Switching & Transmission Technology Lab.(ETRI)
2007 : Ph.D. degree in Computer Science and Engineering, Ewha Womans University
2009~2010 : postdoctoral researcher of the department of computer science, North Carolina State University
2011~present : School of Computer Engineering & Applied Mathematics, Hankyong National University.